Response to Peer Review Comments on the
Draft Plan for the Assessment of Detection and Quantitation Limits
under Section 304(h) of the Clean Water Act

1a. Evaluate the conceptual soundness of the assessment approach presented in the study plan. Is the list of issues included in Item [Event] 3 of the study plan sufficient? Are there other significant issues that EPA should add or delete?

Reviewer A: The assessment approach is conceptually sound. The list of issues in Item [Event] 3 is sufficient. The plan itself seems comprehensive and well thought out. All important issues will be considered in the process. Of course, the most important proof of the soundness of the plan for assessment will be the product of the assessment to follow. At this stage, I believe that EPA has provided a procedure that can address all of the facets necessary to produce a scientifically and statistically valid re-examination of the issue of lower limits for analytical measurements.

Reviewer C: Events 3 through 8 give chief steps in the assessment approach which is proposed. These seem to thoroughly cover many relevant issues. Key to the assessment plan is Event 3, since so many important issues are identified. Giving adequate attention to these issues up front should help to insure that the key issues are understood so that careful choices can be made before the evaluation and selection begins. I can't think of anything to add to the list for Event 3. At the same time, I'd be hesitant to delete anything at this point, although I think that it will be the case that some issues will be found to be not nearly as complex as other issues, and that these simpler issues can be addressed somewhat quickly.

Reviewer E: I think the Plan is sufficient [and . . . see 1b]

Response: Thank you. EPA appreciates the positive feedback that all three of these reviewers provided on the study plan

Reviewer F: Conceptual soundness cannot be evaluated due to insufficient detail. No description is provided on what parameters and/or criteria will be used to assess methods for identifying detection limit or quantitation limit. If one technique/method is to be better than another, what makes it better?

Response: EPA sought input on the soundness of the Agency's plan for identifying criteria that will be used to assess detection and quantitation limit concepts, not on the criteria themselves. EPA intends to develop such criteria after identifying and fully considering relevant issues. EPA believes that establishing criteria before such an evaluation would be premature, and in the later words of this reviewer (see Question 5) would create "the specter that EPA has simply selected criteria that support the conclusion(s) they wish to make." EPA would like to add that the Agency's entire assessment process, including the plan, the identification of issues and appropriate criteria, and the evaluation of various concepts against these criteria will be subjected to formal peer review prior to proposal. (See Events 9 and 10 of the plan.)

Action/Revision Implemented based on Peer Reviewer Response to Question 1a: Revise plan to include a statement clarifying that it would be premature to identify evaluation criteria in the plan itself because issues that might affect these criteria have not yet been fully identified.

1b. How should EPA prioritize the issues [included in Event 3 of the study plan and suggested by commenters]?

Reviewer A: I have some comments below on what I think to be the most significant, high-priority issues.

The entire issue of lower limits of analytical methods is one that has been fraught with intellectual confusion, even when it has not been politically and economically controversial. The term "detection limit" has been used for a number of dissimilar concepts, and this has led to further confusion. For the purpose of this review, I will use a version of Currie's terminology.

The *Critical Level* is an observed value of the response (such as peak area) or equivalently, of the estimated concentration which would occur by chance in a sample in which the analyte was not present only a specified fraction of the time, say 1%. As with any such limit or level, it can depend on the entire analytical method, but is a well defined construct. This is the limit that tells the analyst when detection has occurred. Such detection is almost always quantitative, so long as there is a measured value and a standard error.

The *Minimum Detectable Quantity* is the smallest true concentration of the analyte such that with specified probability, such as 99%, the estimated concentration after measurement will exceed the critical level. This level has no interpretation once measurement has occurred. It is useful for planning, but not for interpretation of existing measurements.

For example, if the standard deviation at zero or a low level of an analyte is 2 ppt exactly, then the critical level at 99% confidence is, using the normal percentage point, (2.326)(2) = 4.652. If the standard deviation is estimated, say by a sample of 7 replicates, then one would use the t distribution with 6 degrees of freedom so that the critical level would be (3.143)(2) = 6.286. Using the latter value for illustration, the practical use of the critical level of about 6 ppt is that any measured concentration exceeding that demonstrates that the analyte is present in non-zero quantities. If we have a measured value of 8 ppt, for example, then the standard deviation (assuming that it is constant over the range) is 2 ppt, so a 95% confidence interval for the true concentration is (using the t distribution again) is 8 ppt ± (2.447)(2)ppt or approximately (3ppt, 13ppt). This is a quantitative assessment. We know the true concentration to within 5 ppt, which is pretty accurate. The coefficient of variation (2/8 = 25%) is not a relevant measure as to whether we have a quantitative measurement of the concentration.

If we use 99%/1% also for the minimum detectable quantity, and if we assume that within the range up to this level the standard deviation of measurement is approximately constant, then the minimum detectable quantity is twice the critical level. In the case illustrated above, this would be twice 6.286, or about 11ppt. This value of 11ppt might be a good spike concentration for quality control, because it should almost always generate a detection signal. However, it provides no information useful in interpreting a measured value of, say, 8 ppt.

Glaser et al. (1981), along with many others, have seriously confused the issue. Consider the oft quoted definition of the MDL: "*…the minimum concentration of a substance that can be identified, measured, and reported with 99% confidence that the analyte concentration is greater than zero…"* The first half of this definition refers to an actual concentration, like the minimum detectable quantity, but the second half treats it as a measured value like the critical level. Clearly, we understand this better than we did in 1981. What the MDL estimates is the critical level, which is the threshold of detection, and it therefore should not be described as an actual concentration.

EPA should use the review that they are engaged in to clarify these issues, and to provide lower limits that are scientifically and statistically valid. In the first place, this means distinguishing carefully between the critical level and the minimum detectable quantity, and specifying when each one is to be used.

Another issue that deserves a hard look and a scientific analysis is that of the so-called limit of quantitation. For an analytical measurement to be quantitative, there are two requirements. First, there needs to be an estimated quantity produced by a known analytical procedure. Second, there needs to be an estimate of the precision of this measurement. Modern metrological techniques (Rocke and Lorenzato 1995; Zorn, Gibbons, and Sonzogni 1997) can provide the precision estimates at all measured values of an analyte. Thus, any measured value by a calibrated system is quantitative, and there is no such thing as a limit of quantitation if it is defined by standard deviation, coefficient of variation and the like. The only scientifically valid definition is akin to the original ML of 1984. If the instrument cannot produce a number, then we are below the limit of quantitation. This will usually be because the peak cannot be identified that should be integrated to produce the estimated concentration.

The procedures provided by EPA and the questions to be examined will be sufficient to deal with these and other issues, so long as the attention is focused on scientifically and statistically valid methods. This review provides an opportunity for EPA to clarify these vexed issues, and provide unambiguous, practical, interpretable, and valid methods for lower limits of analytical methods.

Response: The reviewer has suggested the Agency's prioritization of issues include (1) a clarification of confusion regarding existing detection limit concepts and terminology, including the critical level and the minimum detectable quantity, (2) identification of "lower limits that are scientifically and statistically valid", and (3) examination and scientific analysis of quantitation limits, with the intent of providing "unambiguous, practical, interpretable, and valid methods for [characterizing] lower limits of analytical methods." The reviewer has further commented on the plan that the procedures provided by EPA and the questions to be examined will be sufficient to deal with these and other issues, provided that attention is focused on scientifically and statistically valid methods.

EPA appreciates these comments. EPA agrees that the numerous concepts and definitions proposed by a number of different individuals and organizations have resulted in widespread confusion, misinterpretation, and misapplication.

EPA's primary goal in performing this review is to identify a scientifically valid concept that best meets the Agency's needs under the Clean Water Act. In doing so, EPA will attempt to

address and clarify the confusion that exists among historical concepts and definitions. EPA also will prioritize this reviewer's recommendations concerning the need for statistically and scientifically valid lower limits and unambiguous, practical, interpretable, and valid methods for characterizing these limits.

Reviewer C: With regard to prioritizing the issues listed for Event 3, items a [concepts of the lower limit to measurement], b [concepts in relation to Office of Water applications], f [criteria for design of detection and quantification studies, including selection of concentration levels], and o [accepting the procedures of consensus organizations] seem to be important and may also require more time than some of the other items. However I suspect that at least some parts of the literature review are well underway (or complete). Still, it will be very important to take the results of the literature review and carefully organize and summarize the material (as opposed to just providing a lengthy listing of issues and proposed methods), and this may prove to be a rather difficult job. Item f deals with the experimental design of the studies that will be performed to test various methods. This seems to be a very important part of the overall plan, since the design should take into account many important issues that are identified in completing the many other items of Event 3. Overall, it seems that a large-scale and carefully performed experimental comparison of a good number of well selected methods will be how one finds out which methods are the most accurate. Since time and resource constraints always tend to limit the size of studies that are to be performed, it is important to use a good design in order to produce the most effective results. Items b [concepts in relation to Office of Water applications] and o [accepting the procedures of consensus organizations] touch upon legal and regulatory concerns, and thus may be somewhat complex.

Important items which might not be so time consuming include items e and i. Item e deals with the development of detection/quantitation procedures based on statistical tolerance and prediction. Item i, which calls for consideration of false positives and false negatives, shouldn't take much time, but it is of the upmost importance since the definition of MDL utilizes the probability of a false positive. False negatives are also important, since they are cases where contaminants are present, but a nondetect is recorded. Since false results are such a key issue, it will be important that they are carefully considered when the experimental design is developed. For example, if the false positive rate is to be controlled to be no greater than 0.01, then a large number of trials under the null hypothesis condition of no analyte present must be performed to accurately assess whether or not the false positive rate is adequately controlled. Also, one needs to carefully consider the various ways in which hypothesis testing can be used in the assessment. If the null hypothesis is taken to be $p <= 0.01$, where p is the probability of a false positive resulting from a particular MDL method, then one has to keep in mind that a failure to reject does not provide statistically significant evidence that $p <= 0.01$, since a failure to reject could be due to low power, resulting from too few experimental trials. To truly protect against a method having a false positive rate exceeding 0.01, one may wish to make $p > 0.01$ the null hypothesis and $p <= 0.01$ the alternate hypothesis. It can get to be a bit confusing, since there are two different null hypothesis situations (at two different levels): the null situation of no analyte present, for which we are concerned about the false positive rate of an MDL method, and the null hypothesis that when no analyte is present, the false positive rate is greater than 0.01, in which case a type I error would be a claim that the rate is less than or equal to 0.01 when in fact it is not.

While I do not think that items c [criteria for the selection and appropriate use of statistical models], d [methodology for parameter estimation], and m [outliers] are by themselves unimportant, I do think that it's important to view them as being linked, and to keep in mind that the main concern is to determine which methods for determining the MDL have the greatest accuracy, and not which methods and models are best for a particular stage in the overall determination of the MDL. This is particularly true if the model upon which a method is based is only a loose approximation of reality, in which case the interpretation of parameters may be viewed as being rather fuzzy. In the end, it's whether or not a particular method is reasonably accurate --- viewing the performance of the overall method, and not examining too closely each of it's components. Rather than spend so much time on model selection and assessment, and parameter estimation, it may be better to put more resources into the testing stage, and thoroughly test a large number of methods which have been proposed by others. However, I do think that some thought can go into parameter estimation. For example, if one is uncertain about how to identify and handle outliers, one could replace the commonly used sample standard deviation (it's used in Method 1631B) with a more robust measure of scale. But again, it's important to keep in mind the overall goal of finding an accurate way to determine the MDL. It's not so important to know whether or not a robust measure of scale is a better estimator in the traditional sense of estimating a true distribution parameter accurately. Rather, what is important is finding out whether or not replacing the sample standard deviation with a robust measure improves the performance of the overall determination of the MDL. Fortunately, in some cases, using alternative estimators, in addition to the one originally proposed for use with a particular method, won't require that additional observations be made, but will only require that alternate determinations of the MDL be made using the same set of data.

Response: EPA appreciates the reviewer's comments and recommendations concerning the prioritization of issues and the suggestion that statistical issues be viewed as being linked with the other criteria. EPA will integrate the suggestions into the issue paper and evaluation criteria.

Reviewer E: [I] would prioritize the need for MDLs that are performance driven by the labs. EPA only provides some guidance with ranges of MDLs in a method.

Response: EPA agrees that detection limit capabilities may vary among laboratories. EPA believes, however, that many factors, including laboratory capability, may drive MDL performance. These factors may include, but are not limited to, variations in instrumentation, technological advances, analyst and/or laboratory experience, and even regulatory needs for lower levels of detection. EPA will address this issue when developing its issue paper.

Action/Revision Implemented based on Comment 1b: Include a discussion of existing concepts and definitions in the Agency's issue paper. Prioritize the need for:

- Statistically and scientifically valid lower limits (Reviewer A)
- Unambiguous, practical, interpretable, and valid methods for characterizing these limits (Reviewer A)
- Concepts of the lower limit of measurement (Reviewer C)
- Criteria for the design of detection and quantification studies, including selection of concentration levels (Reviewer C)

- Accepting procedures of consensus standards organizations, which may be complex because of legal and regulatory concerns (Reviewer C)
- Concepts in relation to Office of Water applications, which may be complex because of legal and regulatory concern (Reviewer C)
- Statistical and tolerance levels, which are important issues, but probably not very time-consuming (Reviewer C)
- Consideration of false positives and false negatives, which the reviewer notes will likely require little time but is of the upmost importance and offers specific issues to consider with respect to this item (Reviewer C)
- Performance-driven detection limit concepts (Reviewer E)

EPA also will incorporate Reviewer C's comment concerning the linked nature of statistical issues into its assessment process.

2. Based on your understanding of the issues, are there any types of criteria that you suggest EPA include to evaluate detection and quantitation procedures [Event 4, Criteria Development]? If yes, please identify the criterion and provide a rationale for its inclusion.

Reviewer A: Here are some of what I think to be the most important criteria for lower limits of analytical procedures:

a) Definitions should be consistent, scientifically meaningful, and statistically valid. For some specific comments on this issue, see below.
b) Procedures that implement the definitions should be clear and well defined. They should also implement the definition.
c) The data used to establish specific limits should be realistic. For example, if matrix effects are presumed to be important, then matrix spikes rather than reagent water spikes should be used.
d) There should be no artificial limits. Each defined and implemented limit should be based on a scientific and statistical analysis of the true limits of knowledge, not on preferences of involved parties for higher or lower levels. If we can measure something and provide an estimated standard deviation, then we should be able to use the data in science and in regulation.

Response: Reviewer A has provided criteria that largely reflect the reviewer's prioritization of issues in response to question 1B. EPA agrees these are reasonable criteria and will include them
in the list of criteria against which to evaluate detection and quantitation limit concepts.

Reviewer C: I think that the main focus should be the probability of a false positive result in the null case of no analyte present. This probability should be as close as possible to the allowable limit (say 0.01) without going over the limit. While exceeding the limit is particularly bad, since false detections could result in lawsuits and also contribute to a lack of confidence in water quality studies, being too far under the limit, which results from MDL determinations that are larger than they need to be, hurts the power of the method to detect the analyte when it is

present. Alternatively, and perhaps nearly equivalently, one can view the situation in terms of both false positives and false negatives, with a goal being to find the method with the smallest false negative rate that also satisfies the false positive criterion. The trouble with this approach is that the false negative rate will almost surely depend on how much analyte is present, and one method may be better than another at low levels while the opposite is true at higher levels. (Putting it in statistical terms, there may not be a uniformly most powerful test.)

It should be kept in mind that the methods will be used in a variety of settings, and so the null situation of no analyte present should be tested with a variety of water sources. If one observes differing false positive rates for the variety of water sources (all having no analyte present), then rather than average all of the observed false positive rates, I think it may be prudent to focus on the highest value and require that it satisfy the false positive criterion, since one will want to believe that the false positive rate limit is adhered to no matter what the circumstances (within reason) may be.

Response: EPA agrees that the criteria must reflect concerns about false positives and false negatives. However, the Settlement Agreement schedule for this evaluation does not allow EPA to implement the reviewer's suggestion to test the null hypothesis in a variety of water sources (all having no analyte present). Indeed, EPA is not sure how such tests could reasonably be conducted unless the variety of water sources tested were first manipulated to remove of all possible traces of the pollutant of interest from the matrices in order to ensure that any positive results obtained are, indeed, false positives. The problem with such an approach, however, is that removing the target pollutant from the matrices could also remove any matrix interferences.

Reviewer C: In addition to focusing on the false positive rate, which is tied to the definition of the MDL, I also think that the MDL values from the various methods investigated should be compared to an estimate of the 99th percentile of the null distribution, by which I mean the distribution of determinations obtained in the case of no analyte present. It seems to me that the 99th percentile of this distribution could be viewed as a value for the MDL. Using any value less than this 99th percentile as the MDL would mean that if no analyte were present, the probability of obtaining a measured concentration greater than the MDL would exceed 0.01, and thus violate Glaser's definition (which is also the definition given in Appendix B to Part 136). To estimate the 99th percentile of the null distribution, a simple nonparametric quantile estimator could be used (with the choice of estimator being dependent on the shape of the distribution and the number of observations (where here each observation is a determination of the concentration of analyte from a sample in which no analyte is present)). It should be kept in mind that in order to get a decent estimate of the 99th percentile of a distribution, it may take several hundred observations. So this method of determining the MDL may be impractical for ordinary use. But perhaps it can be used as a spot check in at least a small number of cases.

Response: EPA agrees with the reviewer that obtaining a decent estimate of the 99th percentile of a distribution may take several hundred observations and would be impractical for ordinary use. For that reason alone, EPA is not likely to establish such a test as a criterion for selecting an appropriate detection limit concept. If, however, EPA is able to obtain an existing data set that would allow for such an investigation during the course of its detection/quantitation limit assessment, EPA would attempt to perform the analyses recommended by this reviewer. At this time, EPA is not aware of any detection limit studies containing hundreds of analyses conducted

on samples that were known to be free of the target pollutant of interest yet retained matrix effects.

Reviewer C: I think claims made in the literature about various methods [for establishing detection/quantitation limit concepts] should not be used to judge them, since previous investigations of accuracy were not done the same way for all of the methods. Instead, the determination of the most reliable methods should be based on the results of a large well-designed and properly executed experiment (or series of experiments).

Response: Again, to the extent that EPA can use existing data to evaluate the various concepts, EPA will do so. Unfortunately, EPA has found that very few data sets exist that would allow for such assessments. In part, this is because historical studies typically were performed with very few concentrations, very few replicates at each concentration, or concentrations that were not in the region of interest (i.e., the region of detection). Compounding the problem (and adding to the confusion), is the fact that many of the claims in the published literature are based on information extrapolated from data sets that were not appropriately suited to support these claims. We interpret the reviewer's comment to mean that s/he shares this concern. At this time, EPA is aware of only one data set that includes multiple replicates of varying known concentrations above, at, and below the estimated detection limit. Even this data set, which was collected by EPA in 1998-1999 as part of its detection/quantitation limit assessments, is limited by the fact that it involves only single-laboratory determinations. If data, time, and resources become available for such an assessment prior to proposal, EPA will be happy to implement this reviewer's suggestion.

Reviewer C: Finally, with regard to criteria, some weight has to be put on simplicity and ease of use. The method(s) selected will be used by a large number of people, and some will be more skilled than others.

Response*:* EPA agrees with the reviewer's comment that the criteria should give weight to simplicity and ease of use by a large number of people with varying skills.

Reviewer E: I do not have any criteria in mind.

Response: No response required.

Reviewer F: EPA has not provided a single criterion for performing this evaluation. Without some idea of what they are planning to do, it is not possible to suggest alternatives.

Response: As noted previously, EPA does not believe it is appropriate to specify criteria before fully assessing related issues. EPA does, however, intend to seek formal peer review on its selected criteria at a later date. The purpose of *this* peer review was to seek comment on EPA's plan for identifying issues, developing criteria that reflect these issues, and evaluating concepts against those criteria. EPA also sought to use this peer review to seek preliminary recommendations on issues, criteria, and concepts that EPA should not overlook when implementing the plan.

Action/Revision Implemented based on Peer Reviewer Comments on Item 2: EPA adopted each of the criteria recommended by the peer reviewers into its final selection of criteria as follows:

Reviewer	Reviewer's Recommended Criteria	Maps to EPA Criterion Number(s)
A	A - Consistent, scientifically meaningful and statistically valid	1
	B- Clear and well-defined procedures that allow for implementation of the definition	1 and 3
	C - Rely on realistic data, such as matrix samples when matrix issues are a concern	3
	D - Establish limits on scientific and statistical analysis, not on personal or organizational preferences for higher or lower limits	1, 2 and 4
C	Criteria should address appropriate risk of false positives and false negatives	4
	Criteria should recognize the need for simplicity and ease of use by a large number of people of varying skills	3

3. What existing detection and quantitation procedures should EPA evaluate or focus on? Procedures from the open scientific literature? Voluntary consensus standards procedures? Procedures from other organizations?

Reviewer A: I believe that the EPA should take a fresh look at all of the suggested definitions and procedures It is correct to evaluate suggestions from the scientific literature, from standards organizations, and from other organizations. However, it may be that none of these procedures meets the criteria of 2) [i.e., criteria suggested in commenter's response to charge question 2], and meeting those criteria is far more important than precedent.

Response: EPA appreciates this feedback. EPA will attempt to evaluate each unique concept or definition available in the literature. (EPA believes that there is significant overlap in concept among articles describing different detection limit terms.) EPA agrees with the reviewer's comment about the possibility that none of the existing definitions and procedures will meet the selected criteria and that meeting those criteria is the primary objective of the study.

Reviewer C: Of course, methods currently used by the EPA, such as Method 1631B, should be included. In addition, I'm in favor of including several promising methods that are to be identified from the open scientific literature, as well as perhaps a small number of alternate procedures that are used by other organizations, provided that some justification can be found for them other than that they are sometimes used. (I fear that sometimes a method can become popular for no real good reason. One organization may select a method for use somewhat haphazardly, and then another organization can pick up the method just because others have used it. Since methods are sometimes selected based in part on political considerations, or concerns about practicality that may not be pertinent in all cases, I tend not to give a lot of weight to the

popularity of methods, and I'm much more impressed by studies, performed by people other than the originators of the method, that demonstrate that a method is reasonably accurate.)

Methods from the literature are nice to consider, because as time passes, refinements seem to be made upon previously proposed methods, and one can be hopeful that the class of methods that get the most attention of this sort are the ones that perform best, and thus hope that recent papers giving the latest adjustments to an established general scheme will include some that are (near) state-of-the-art. Of course, one has to be wary of authors who are too eager to publish and develop a small adjustment in hope of getting a quick and easy publication. Also, since many published methods seem to be highly derivative of previously published ones, there might not be enough variety found in the literature and so one may want to consider other sources.

Response: EPA believes that this reviewer may have misunderstood the Agency's question. EPA intended to seek input on procedures for determining detection and quantitation limits, not on measurement methods for specific analytes.

Reviewer C: With regard to identifying alternate methods from the literature review, I think that the 1998 memorandum from Raphael Kuznetsovski of SAIC should be of great help. There are 27 papers and books described in the method detection limits and quantification section. The summaries provided for each of the 27 works can be used to narrow the focus to perhaps a dozen or fewer key entries worthy of careful consideration. In general, I recommend examining recent articles, slightly older articles from the chemistry literature by authors who are well established in the field, and also articles from the statistical literature in order to perhaps gain a slightly different perspective.

Response: EPA agrees that the 1998 literature review assembled by SAIC under contract to EPA will be of help in identifying pre-1998 methods from the published literature. EPA has supplemented these articles with a literature search to identify post-1998 articles that may be relevant to this evaluation. As noted above, EPA agrees that many of the articles can be consolidated into a smaller subset of unique concepts or procedures.

Reviewer C: I'll now make some brief comments about some of the works described in the SAIC memo. (Since complete citations are provided in that memo, I won't duplicate such information here.) Several of the papers pertain to the Hubaux and Vos approach: the original 1970 paper, the 1978 paper by Bayley, Cox, and Springer, and the 1997 paper by Coleman, Auses, and Grams, with this last paper being by far the most recent of the three, and perhaps worthy of a close look. Gibbons has written a lot of pertinent articles, and I think several of his works should be studied, particularly those which compare various methods, like his 1994 book. Similarly, Currie has written a lot, but some of his work is rather dated, and some of it deals with related issues and not particular methods. But his 1997 paper may be of interest, since it touches on assumptions and approximations. Davis' 1994 article, the 1991 paper of Lambert, Peterson, and Terpenning, and the 1997 paper of Spiegelman and Tarlow are examples of works from the statistical literature that could be of interest, and the 1987 paper of Clayton, Hines, and Elkins and the 1988 paper of Garner and Robertson are papers from the chemical literature that could be considered, with the later paper of particular interest since it touches on the issue of assumptions.

Response: EPA plans to consider the specific papers cited by this reviewer in its evaluation.

Reviewer C: Finally, I think that it will be important to consider variety with regard to the complexity of the methods to be chosen. It might be good to include some simple methods in addition to those with are rather involved. Not only are the more complex methods more difficult for some to accurately use, but it may be the case that some of the simpler methods could outperform some of the complex ones, due to poor assumptions underlying some of the methods.

Response: EPA agrees that concepts relying on simple procedures may outperform those involving complex ones, and that the Agency's assessment should include an evaluation of both simple and complex methodologies.

Reviewer E: I think EPA should focus on the MDL approach and the detection limit should be performance driven. Each lab should develop their own MDLs and the EPA method should provide some examples and some expectations for each analytes and matrix but let the labs demonstrate their own MDLs. The "market" will then drive the MDLs. If there is a need for more sensitive MDLs then the successful labs will improve their MDLs.

The "environmental chemistry" community has been using the Fed. Reg. MDL method for about 10 years to demonstrate detection limits. I feel this is an acceptable approach.

Response: While EPA agrees with the reviewer's suggestion that allowing the market to drive detection limits would be beneficial, the Agency does not believe that it is appropriate to limit its assessment to the MDL procedure. Doing so would violate the spirit of the Settlement Agreement.

Reviewer E: Another decision EPA needed to make is to define the factor that is used to convert the MDL to a Reporting Limit, or what every it will be called. For example if the RL or ML is defined as 3.18 times the MDL then everyone could understand and use it even though they may not agree.

Response: While EPA does not necessarily agree (at least without further evaluation) that reporting limits should be based on a simple multiplier of the MDL, EPA agrees that the issue of reporting limits differs from detection limit concepts and will consider the use of multipliers and other suggested approaches. Given the limited schedule and resources available, however, EPA will necessarily limit any discussion of reporting limits to their context within Clean Water Act programs.

Reviewer F: I believe the charge to EPA is that they look at all existing procedures, see if they can come up with any new ones and then, with some predefined rationale, pick the best one. This sounds like a justifiable charge. If the intent of this question is to determine if I have a favorite that I want to be sure is on the list, I do not.

Response: EPA believes that significant overlap among the existing procedures allows for classification into broad groups, much as described by Reviewer C above. Given the limited schedule, EPA will attempt to review as many concepts as possible, including each of those recommended by the Peer Reviewers and any other concepts that the Agency identifies as being likely to meet the selected evaluation criteria.

Action/Revision Implemented based on Peer Review Comments to Question 3: Evaluate concepts from the published literature against the selected criteria that represent a range of complexity, from simple to complex. At a minimum, include the following papers:

- Hubaux and Vos approach (the original 1970 paper, the 1978 paper by Bayley, Cox, and Springer, and the 1997 paper by Coleman, Auses, and Grams
- The Gibbons papers, including his 1994 book that compares various methods
- Currie's work, particularly his 1997 paper
- Davis' 1994 article (statistics)
- Lambert, Peterson, and Terpenning's 1991 article (statistics)
- Spiegelman and Tarlow's 1997 paper (statistics)
- Clayton, Hines, and Elkins 1987 paper (chemistry)
- Garner and Robertson 1988 paper (chemistry)
- EPA's method detection limit and similar concepts
- Literature concerning detection limit multipliers or other approaches to derive acceptable reporting limits

4. Are the data sets listed in the plan appropriate for use in the detection and quantitation reassessment? Are there other data sets that you recommend EPA evaluate?

Reviewer A: I am familiar with some of the data sets proposed, but not all of them. Data sets for a procedure like this one should have several to many replicates at a number of concentrations from zero or near zero, to near the critical level to high concentrations.

Response: EPA agrees.

Reviewer C: Based on my rather limited knowledge about the data sets referred to, they do seem appropriate for use. At this time, I don't have any other data sets to recommend. However, it may be good to produce some new data for this important study, since existing data may not be ideal.

Response: Unfortunately, EPA does not have the time or resources needed to produce new data to support this evaluation. EPA will continue to consider any new data made available to the Agency for this use, however.

Reviewer E*:* I am aware of MDL data sets for the MDLs in method 1638 that could be evaluated.

Response: MDL data sets will generally not meet the needs of this project. As pointed out by Reviewer A, data sets for a procedure like this one should have several to many replicates at a number of concentrations from zero or near zero, to near the critical level to high concentrations.

Reviewer F: The Plan indicates that Episode 6000 data will be reviewed. Are these the only data that will be reviewed? Are these data associated with Method 1631B? Are there other data that could be used in addition? If so, will they be used or excluded?

There is only one data set identified in the Plan, Episode 6000 data. If these are mercury data derived from Method 1631B, they are appropriate for that method. However, because I haven't seen them, I don't know. However, I would be surprised to hear that Episode 6000 data are applicable to all EPA methods. If they are, how was this determined? I know EPA has other data associated detection studies because I've seen EPA statisticians present them at various technical meetings. Also, based on the references identified in the memo to Chuck White from Raphael Kuznetsovski, it would appear that there are a number of other data sets outside EPA. Perhaps the authors would be willing to share them. What I'm not sure of is whether or not the Plan is intended to address other EPA methods.

Response: EPA agrees that the Episode 6000 data set was not adequately described in the draft plan. The Episode 6000 data set consists of a series of iterative MDL studies performed on 11 measurement techniques, including ICP/AES, ICP/MS, GC/MS, and classical wet chemistry. EPA also agrees that additional data should be reviewed as part of this assessment, provided that those data are from studies directed at characterizing variability in the region of detection. EPA will revise the plan to include an improved description of the Episode 6000 data set and of the Agency's strategies for identifying and selecting other data to be included in the study.

Action/Revision Implemented based on Peer Review Comments to Question 4: Revise the draft plan to clarify the nature of the Episode 6000 data set and the Agency's strategy for identifying and selecting additional data sets. Also, we should solicit additional data sets in the proposal.

5. Is it appropriate to include specific selection criteria in the plan or is it more appropriate to develop selection criteria based on the analysis of issues that will be conducted as part of the plan?

Reviewer A: It is better to develop selection criteria during the development of the project. However, possible criteria should be included in the plan.

Response: EPA is pleased that the reviewer agrees it is better to develop selection criteria during project development. With respect to documenting possible criteria in the plan, EPA believes that modifying the plan to include criteria that are selected at a much later date (i.e., after fully considering the issues) will cause confusion. Therefore, EPA will document the selected criteria separately.

Reviewer C: I believe that it is better to firm up the selection criteria as the study progresses, thinking that people will develop a better understanding of the issues as they put more thought into them. Plus, instead of specifying the criteria up front, it may be better to wait and see what happens and then examine the pros and cons of the various methods when making specific comparisons.

Response: EPA is pleased that this reviewer also agrees it is better to develop selection criteria during project development.

Reviewer E: I think the selection criteria should be included so each lab demonstrates their MDLs by analysis of replicate samples of actual field samples , either spiked or unspiked depending on the concentration in the matrix. EPA would provide the guidance in terms of number of replicates

Response: In developing it's selection criteria, EPA will consider this reviewer's preference for criteria that include performance in real-world matrices. The existing MDL procedure provides guidance on application to matrices other than reagent water; e.g., to field sample matrices. EPA may produce a separate guidance document to clarify the detection/quantitation procedures ultimately selected.

Reviewer F: It is not just "appropriate" to include specific selection criteria, it is essential if an objective assessment is to be conducted. Otherwise there is the specter that EPA has simply selected criteria that support the conclusion(s) that they wish to make. This is likely to be obvious to the plaintiffs in the suit against EPA.

Response: EPA disagrees. EPA believes that it is not appropriate to select criteria without a rationale, and the rationales for various criteria cannot be fully developed without a full evaluation of the related issues. Other arguments against defining the criteria in the planning stage were provided by Reviewer C (see above).

Action/Revision Implemented based on Peer Review Comments to Question 5: None.

V. OTHER COMMENTS

A. Summary-Level Comments

Reviewer C: I think the plan is very thorough, with the various events called for touching on many important issues, and I can't find anything that seems to have been overlooked. The issues considered deal with the legal aspects of the situation, and also with doing good science and searching for a reliable MDL determination procedure.

Response: EPA appreciates the positive feedback on the plan.

Reviewer F: The Plan is too vague and lacking in technical detail to generate an assessment of its scientific merit. The approach described in the Plan does not contain sufficient detail to determine whether or not it is valid. Neither does it contain enough information with respect to how the assessment will be done to determine the likelihood that it will be objective and/or useful.

Response: EPA is disappointed that Reviewer F finds the plan too vague, yet Reviewer F offers no specific details to enhance the plan.

Action/Revision Implemented based on Summary Level Peer Review Comments: None.

C. Concern that Plan Does Not Meet Requirements of Clause 6 of the Settlement Agreement

Reviewer F: This reviewer has read and reread these documents. The second paragraph of the Charge to Peer Review Panel requests an opinion regarding the validity of the Plan and the likelihood that it will result in an "objective and useful assessment of detection (limit) and quantitation (limit) procedures". None of the five Questions to Peer Reviewers (included in the Charge to Peer Reviewers) address the purpose indicated in the Plan's introduction as the reason the plan was developed, namely, responding to Clause 6 of the settlement agreement. In addition, only the Plan document identifies Method 1631B; the other documents seem to address EPA methods in general. Consequently, I will try to address all the review considerations starting with the Plan.

The Plan identifies 4 key items associated with satisfying the requirements of Clause 6 of the settlement agreement. These are: Reassessment of EPA's existing MDL and ML procedures; External review of the reassessment; External review of any alternate procedures EPA has under consideration; and Proposal of revised/alternate procedures, if warranted. This is to be accomplished by the end of February 2003. To this end, EPA has established a series of tasks and milestones leading up to the due date. Although not specifically stated, it is presumed that a document will be generated which addresses the four items in Clause 6.

Response: The study plan contains a section near the end titled, "Relationship of the Plan to the Settlement Agreement," that specifies which plan events will meet the requirements of each subpart of Clause 6 of the Settlement Agreement. EPA believes this section is sufficiently clear, but in response to Reviewer F's concerns, will also integrate these linkages directly into the plan where possible. With respect to the reviewer's question concerning documentation, EPA intends to develop one or more documents to detail the Agency's implementation of this plan. This documentation will include the Agency's identification and assessment of issues, the criteria selected for evaluating various concepts, the evaluation of various concepts, and the rationale for the final selection. EPA expects to incorporate each of these documents into a Technical Development Document that will support the FR notice.

Reviewer F: At the end of the Introduction to the Plan, there is a statement that the "document contains the detailed plan for the reassessment called for in Clause 6 …". In fact the Plan contains very little detail. The only thing that comes close to detail is the listing of "issues" that are part of Task 3. There is no description of the processes and/or criteria that will be used in the reassessment. EPA should identify the specific criteria that will be used to select the most appropriate mechanism(s) for determining detection limit(s) and/or quantitation limit(s). If the criteria are not identified up front, it may appear that EPA first selected mechanisms, then created justification to support them. This is what Clause 6 seems to be trying to avoid.

Response: EPA disagrees for reasons explained in responses to Issues 2 and 5 in this document.

Reviewer F: Presumably the people involved have the skills and knowledge, and EPA will allow them the time and provide the resources, needed to do the job. Unfortunately there is no way to assess this based on the Plan. EPA identifies a "Core Workgroup consisting of measurement analysts and statisticians within the Office of Science and Technology" but does

not provide information on their background or experience. The success of this project will depend on the people involved, so this is critical information. In summary, it is very difficult to see how this Plan will result in EPA being where they want to be when they want to be there. They may succeed, but working to this Plan will not be the reason for their success.

Response: The purpose of this peer review was to seek input on EPA's plan for evaluating concepts and methodologies supporting detection and quantitation limit concepts. The plan outlines a series of tasks and a schedule for accomplishing those tasks. EPA will provide the resources and staff necessary to implement this plan. These resources include staff who are experienced and trained in the areas of analytical chemistry, environmental statistics, and implementation of Clean Water Act programs, and are familiar with many of the detection and quantitation limit issues identified in the plan. In addition, materials developed by this team will be reviewed by other individuals throughout the Agency, including those that represent Regional perspective and legal perspectives, as well as those that offer insights from other environmental programs (such as the Safe Drinking Water Act).

Action/Revision Implemented based on Peer Review Comments concerning adherence to the Settlement Agreement: Clarify the linkages between the Settlement Agreement and the planned activities.

D. Unclear whether Plan Applies to All Methods or Only Method 1631B

Reviewer F: It is also not clear whether the Plan is intended to address only Method 1631B or EPA methods in general. If it is EPA methods in general, the magnitude of the job is much, much greater and it seems unlikely that a valid technical effort could be conducted in the time frame identified.

Also, because Method 1631B is cited in the Introduction along with Clause 6, it is assumed that the Plan document is intended to address only this method.

Response: Although the Settlement Agreement was established in response to a legal challenge to Method 1631B, the terms of the Agreement dictate that the Agency evaluate detection and quantitation limit concepts as they apply to all CWA programs. EPA will modify the plan to indicate more clearly that the scope of this effort is to address methods used in CWA programs.

Action/Revision Implemented based on Peer Review Comments concerning Scope: Revise the plan to clarify its scope.

E. Confusion over Milestones/Event Sequence

Reviewer C: In the schedule given on page 12 some events seem out of sequence. From pages 8 and 9 it seems as though Events 5 and 6, which deal with evaluation and reassessment, should follow Event 4, which is criteria development, and this seems sensible to me. Yet on page 12, it appears as though Event 4 lags behind Events 5 and 6.

Response: EPA agrees. EPA has revised the schedule to address these and other concerns.

Reviewer F: EPA has organized this effort into 16 tasks/milestones that are identified in a chart on page 12. The first two tasks have been completed, the resulting output being the documents under review. The organization of the remaining tasks, however, is somewhat confusing. First, there seem to be six first-draft reports, excluding the Federal Register Notice, plus two second-draft reports. I suspect these are sections in the final report document but this is not clear. It appears that the drafts are assembled into a single report as the interim, and only milestone, of Task 9. External Peer Review is then preformed over a three-month period shown as Task 10, starting in May of '02. Final selection of the "Recommend Option" takes place as a milestone (task 9) after External Peer Review is complete.

However, even though the report has been sent out for external peer review, development tasks (Tasks 3, 4 and 7), evaluation tasks (Tasks 5 and 8), and the reassessment task (Task 6), continue to run on through Feb '03. Consequently, it appears that the material developed in these tasks after April '02 will not be subject to external peer review. This appears to be inconsistent with the requirements of Clause 6. Further, why should these tasks continue past the time the Recommended Option has been selected? If the tasks continue does that mean some other Option could replace the one identified in task 11?

In addition, some of the tasks identified in the Plan do not appear to be clearly related to dealing with the four points of Clause 6. The first point of Clause 6 is addressed by Task 6, the second point by Tasks 9 and 10, the third point by Tasks 8, 9 and 10, and the fourth point by Task 7. Tasks 1 and 2 were used to create the documents reviewed here. That leaves Tasks 12 through 16, which appear to be EPA administrative issues that must fit into the time line prior to the completion date. If this is the case, it is difficult to understand how the development, evaluation and reassessment tasks identified above could continue while EPA is writing a Federal Register Notice (Task 12), the Agency is reviewing (Task 13) and OMB is reviewing (Task 14).

Response: EPA agrees and has revised the schedule to address these and other concerns.

Action/Revision Implemented based on Peer Review Comments concerning Schedules:
Include revised schedule in the plan.

F. Comments Supporting Ability to Change MDL Over Time

Reviewer E: [Study Plan, Page 4, Section ii. Method Development and Promulgation] I strongly believe that detection limits should not be cast in stone when the method is promulgated. The DL's often improve radically as labs become more proficient and improvements are made to techniques. I have experienced clients and regulators who flag data as below the DL because it is below the Method-specified DL but well above our achieved DL!

Response: EPA agrees that detection limits are subject to change and will bear this in mind when defining criteria for evaluating concepts. Again, this is an implementation issue that will affect any concepts chosen by EPA. How we ultimately address this issue may not be part of this effort but, as noted earlier, may be addressed as part of a guidance document. Alternatively, we can provide improved guidance for application of whatever procedure goes into 40 CFR 136 in place of the current Appendix B. We will point out that states and others can still do differently.

Reviewer E: [Study Plan, Page 5, Section vii. Descriptive versus Prescriptive Uses of Lower Limits to Measurement] I am very much in agreement with this statement which I assume means that the detection limit can change depending on the measurement method and the proficiency of the lab.

Response: EPA appreciates the positive feedback and will bear this in mind when defining criteria for evaluating concepts.

Reviewer E: [Study Plan, Page 6, Section j. Measurement Quality Over the Life of a Method] I agree with this statement that the measurement quality will usually improve over the life of the method. This is a very important issue and therefore the DL for a method should be allowed to change as the equipment and analyst change.

Response: Thank you for the positive feedback. Again, EPA will bear this in mind when defining criteria for evaluating concepts.

Reviewer E: [General comments] Each EPA method should include a statement about the range of MDLs that have been demonstrated during the validation of the method. Each EPA method should state that any lab using this method must demonstrate their own MDL, which may be higher or lower than the range mentioned in the method.

Response: EPA will consider this recommendation as part of its overall evaluation.

Reviewer E: The labs can inform their "clients" about their lab MDLs and the "environmental chemistry market" will drive these MDLs. If a regulation requires low detection limit then labs will be motivation to demonstrate the needed MDL.

Response: EPA will consider this recommendation when defining criteria for evaluating concepts.

Action/Revision Implemented based on Peer Review Comments concerning the ability of the MDL to Change over Time: When identifying and addressing issues, include Peer Reviewer E's concerns about the need for concepts that allow detection limits to be adjusted according to changes in technology and lab proficiency.

G. Clarification Needed Regarding the Term "Outliers"

Reviewer E: [Study Plan, Page 7, Section m. Outliers] The term "outliers" is confusing as used in the statement. My understanding of outliers are data points that can be separated from the other data points with a statistical test called an "outlier test." There should be an acceptable procedure to determine when data is not appropriate to be included in an MDL.

Response: For the purposes of EPA's evaluation of detection/quantitation limit concepts, EPA will investigate the various definitions of "outlier" and choose the definition most consistent with application to these concepts.

H. Recommend Exploration of Procedure to Develop Theoretical MDLs for Solids Based on Liquid MDLs

Reviewer E: [Study Plan, Page 8, Section t. Cost to Implement Limit Procedures] A procedure for developing "theoretical MDL's" for solids, such as sediment and tissue, based on liquid MDL's should be explored. For example a true MDL for mercury in fish tissue cannot be experimentally determined because all fish tissue has relatively high levels of mercury compare to the sensitivity of the method. However, using water derived MDL it is possible to calculate the MDL for fish. This situation is true for many metals that are naturally abundant in sediment, soil and tissue.

Response: EPA will address the issue of "theoretical MDL's" in its evaluation. EPA would like to note, however, that the Agency has successfully used matrices other than water to simulate performance in tissue. For example, EPA has found that egg and chicken breast can be used to establish 'tissue' detection limit studies for mercury using Method 1631.

I. Typos

Reviewer F: [Study Plan, Page 10] There are a few typos, judgment is misspelled, and Task 16 on page 10 should identify December 2002, not 2001. It's not the nits that need to be addressed, it's organization and the lack of detail.

Response: EPA has corrected the typos on Page 10 of the plan and has revised the milestone/event schedule to correct dates as appropriate.

J. Concern Regarding Errors in Current Appendix B to Part 136

Reviewer C: This doesn't pertain explicitly to the 12 page plan. But in reviewing supplementary material that was sent to me along with the plan, I noticed a few typos in Appendix B to Part 136. The most serious mistake is in part (b) of Step 7. If the ratio of variances is less than 3.05, then one would compute the pooled standard deviation. So the first >3.05 should be <3.05. (One can check this in the original description of the method given in the 1981 paper by Glaser, Foest, McKee, Quane, and Budde.) Less serious are two minor typos: (i) below the sum at the bottom of page 2 it should just be i = 1, not i = 1S; (ii) on the 3rd line of page 3, S should be replaced by a sigma.

Response:. The procedure as published in the code at 40 CFR 136, Appendix B appears to be correct.

K. Concern Regarding Method 136B

Reviewer C: I'll also comment that I'm a bit suspicious of Method 136B. The justification given for the approach in the paper by Glaser et al makes use of assumptions and approximations. Since the final formula given for the MDL doesn't seem quite right, I wonder if perhaps too many assumptions were made along the way. Also, distribution skewness could harm the accuracy of the method.

Response: EPA will consider these concerns when evaluating concepts.